The Application of Professionally Accepted Standards for Reliability and Validity to the Collection of Evidence

Prepared by C. Taylor and J. Willhoft

July 14, 2006

One of the three options that have been legislated as alternatives to performance on the Washington Assessment of Student Learning (WASL) as a means to earning a Certificate of Academic Achievement is a collection of work samples, also referred to as the "Collection of Evidence[1]" (COE). Legislation requires that the guidelines and protocols for submission and the criteria used for scoring "meet professionally accepted standards for a valid and reliable measure of grade level expectations and the essential academic learning requirements." The purpose of this paper is to describe a set of standards that can be applied to the collection of student work samples from the classroom environment that can be considered as meeting those professionally accepted standards.

The process recommended by OSPI to the State Board of Education (SBE) is that the Standards proposed by Taylor and Nolen shown in Tables 1a and 1b and provided in more detail in Appendix A be reviewed and approved by the National Technical Advisory Committee (NTAC). That approval will assure the SBE that the criteria for reliability and validity against which the COE will be judged meet "professionally accepted standards". The review and approval of a set of reliability and validity standards is on the July 28-29 agenda for the NTAC. Once the NTAC adopts a set of reliability and validity standards for the COE, the design features

[1] Collections of Evidence are subject specific (i.e., reading, mathematics, and writing) collections of classroom-based assessments or work samples for individual students that demonstrate comparable curriculum standards as those assessed by WASL.

of the COE will be submitted for their review. The NTAC will be asked to reach consensus on the alignment of design features of the COE that address the standards. That effort will be completed in early August, and will be presented to the SBE at its August meeting.

Two sets of standards have served as source materials for the preparation of this paper. The *Standards for Educational and Psychological Tests* were developed jointly by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). The fourth edition of these standards was published in 1999. This document is widely accepted within the community of measurement professionals as encompassing the standards to be met by assessments that are commercially-developed or are used in large scale public assessment systems. The second document is a paper by Taylor and Nolen (1996) in which the authors adapted the *Standards* document for application to classroom-based assessments, focusing on the 'half dozen or so' ways of gathering evidence for the validity of the inferences made from test scores presented by Messick (1989). This paper also used a subsequent treatment by Taylor and Nolen (2005), which addressed issues of reliability in classroom-based assessments. That paper elaborates on three topics found in the literature regarding the reliability of inferences made from test scores: a) standardization of directions (Cronbach, 1971), b) interjudge agreement, and c) internal consistency (Cronbach, 1971). For a more detailed discussion of the standards for reliability and validity, see Appendix A.

Tables 1a and 1b summarize the six validity and three reliability standards developed by Taylor and Nolen.

Table 1a: Validity Standards for Classroom-based Assessments

Validity Std 1: **Authenticity**	Do the knowledge and skills required by the assessment tools reflect the way people in the subject area or discipline actually use their knowledge and skills?
Validity Std 2: **Cognitive Process Validity**	Do assessments actually draw out the targeted knowledge and skills?
Validity Std 3: **Consistency with other Measures**	Do students perform similarly on different measures of the same knowledge and skills?
Validity Std 4: **Curricular Validity**	Does the assessment fit the content taught and the instructional methods used?
Validity Std 5: **Minimizing Bias**	Does the assessment tool, process, and/or event work equally well across individuals, groups, settings?
Validity Std 6: **Consequences**	Are there negative consequences that could be prevented if assessment tools, processes, events, or decisions had been more valid?

Table 1b: Reliability Standards for Classroom-based Assessments

Reliability Std 1: **Sufficiency**	Is there sufficient evidence of student learning so that one can make a dependable statements about what each student has learned in relation to the learning targets?
Reliability Std 2: **Clarity of Expectations**	Do the test items or performance directions provide clear, unambiguous expectations for students?
Reliability Std 3: **Scoring Quality**	Are the scoring rules and evaluation processes systematic enough to ensure consistent evaluation across students and over time?

Tables 2a through 2g summarize the pertinent design features of the COE that address the validity and reliability standards provided by Taylor and Nolen.

Table 2a:
Protocols – Directions to the COE users to indicate the types of evidence needed for each subject area

Writing Protocol

There are to be 5 to 8 written samples that together demonstrate proficiency in idea/development, organization, style, and the use of conventions resulting in writing work samples that address the EALRs. The collection addresses the range of standards in both breadth and depth, focusing on work samples that demonstrate expository and persuasive prose.

- At least two expository non-timed essays
- At least two persuasive non-timed essays
- Of the 5-8 samples, 3 work samples (including the on-demand sample) may not include any adult assistance beyond setting the prompt and the parameters for an effective paper. Other work samples may include drafts read with teacher input and general comments (e.g., "You need to check for spelling errors." or "You need to rework your conclusion to wrap up your writing and give your reader something to think about.")
- At least one expository or persuasive on-demand essay, timed and supervised in class.

Reading Protocol

There must be 8-12 work samples that together demonstrate understanding and application of reading skills that are directly connected to the EALRs. The collection must address the range of standards in both breadth and depth across the genres of literary and informational texts. The work samples should be "writing in response to reading," in order to demonstrate the use of text-based support as evidence of ability.

- one half of the work samples that can be scored for more than of the three literary strands;
- one half of the work samples that can be scored for more than of the informational strands;
- one work samples that can be scored as a short literary analysis paper that features discussion of a novel, short story, poem, narrative essay, autobiography or biography
- one work sample that can be scored as a short informational paper that features the discussion of a magazine article/newspaper article, of a text book section on historical events, or a textbook description of a scientific process (use of content area coursework outside of language arts is encouraged).
- At least one work sample produced in an "on-demand" setting
- All texts used in the work samples must meet high school expectations for rigor of reading material. The work samples must be comparable in rigor in skill and content to the High School Reading WASL.

Mathematics Protocol

There must be 8 to 12 work samples that together demonstrate understanding and application of the five mathematics content strands and four process strands that are connected to the EALRs. The collection addresses the range of standards in both breadth and depth. The work samples should be "rich problems" in which one or more content strand and one or more process strands are represented.

- Work samples of moderate or high complexity to ensure moderate or high level cognitive demands of the student
- At least two high school level work samples that and can be scored for each target from a strand of EALR 1:
- At least two high school level work samples can be scored for each target* from a strand of EALRs 2 through 5:
- Work samples that combine a content strand from EALR 1 and a process strand from EALRs 2 through 5. Work samples for EALRs 2 through 5 must be distributed across EALR 1 content strands.
- Work samples you select must combine at least one content strand from EALR 1 and at least one process strand from EALRs 2–5.
- One work sample that must be completed in an "on-demand" setting where students are provided an assignment to complete within a class period and without any teacher or peer assistance.

Table 2b: **Sufficiency Review – Process used to determine that all of the WASL learning targets for a domain are included in the collection**
Writing Protocol
In order to meet the sufficiency guidelines for successfully submitting a Writing Collection of Evidence, the student and teacher preparing the collection must comply with the COE guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored.
Reading Protocol
In order to meet the sufficiency guidelines for successfully submitting a Reading Collection of Evidence, the student and teacher preparing the collection must comply with the COE guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored.
Mathematics Protocol
In order to meet the sufficiency guidelines for successfully submitting a Mathematics Collection of Evidence, the student and teacher preparing the collection must comply with the following guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored

Table 2c: **Directions for student assignments**	
Writing Protocol	*In the "Work Sample Documentation Form" teachers must provide documentation that the work sample demonstrates the state standards in writing. For each work sample, students must check one of the first three boxes on the form as well as the type of draft, process, and teacher-assisted for the work samples in the collection. The teacher must check that an "on-demand" essay is present in the collection. In the last box—teacher assistance—the student must describe what type of assistance he/she received beyond setting the prompt and the parameters of an effective paper.*
Reading Protocol	*In the "Work Sample Documentation Form" students and teachers must check all of the learning strands, both literary and informational. The student must provide of the titles of the texts must be provided to check the rigor of the readability of the texts. The student and the teacher must check each work sample to make sure that each sample addresses at least two strands. The student must identify which work sample is the short literary analysis paper and which is the short informational analysis paper. The teacher must check that an "on-demand" essay is present in the collection.*
Mathematics Protocol	*In the "Work Sample Documentation Form" students and teachers must check that all work samples address every high school content strand. Each work sample must address both a content strand and a process strand. Teachers must check that work samples meet the "rich problem" and high school level mathematics expectation. Students must check that each column and row have two entries. There must be an "on-demand" check*

Table 2d:

Scoring rules used to evaluate the collections – Performance criteria for the scoring rubrics used for each collection are given below along with an indication of the subject area EALRs and components within each EALR that are the focus of the performance criteria. Links to the EALRs are keys to authenticity validity.

Writing Criteria

Content, Organization & Style

- Has clear, focused main ideas or positions (EALR 1, Component 1)
- Elaborates by using reasons/arguments supported by well-chosen and specific details, examples, anecdotes, facts and/or statistics as evidence to support ideas or positions (EALR 1, Component 1)
- Includes information that is thoughtful and useful for the audience to know (EALR 1, Component 1)
- Organizes writing to make the best case to explain ideas or support positions (EALR 1, Component 2)
- Composes introductions that draw the reader into the main ideas or positions (EALR 1, Component 2)
- Writes conclusions that leave the reader with something to think about (EALR 1, Component 2)
- Organizes writing into effective, cohesive paragraphs (EALR 1, Component 2)
- Provides transitions which clearly serve to connect ideas (EALR 1, Component 2)
- Uses language effectively by exhibiting word choices that are effective and appropriate for intended audience, purpose, and form (EALR 1, Component 3)
- Writes (where appropriate) sentences or phrases that are varied in length and structure (EALR 1, Component 4)
- Provides the reader with a sense of the person behind the words (EALR 1, Component 5)

Conventions

- spelling of commonly used words (EALR 1, Component 6)
- capitalization (EALR 1, Component 6)
- punctuation (EALR 1, Component 6)
- exhibits the use of complete sentences except where purposeful phrases or clauses are used for effect (EALR 1, Component 6)
- indicates paragraphs consistently (EALR 1, Component 6)

Reading Criteria

Comprehension of main ideas and details of literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text

- Identifies the main theme/main idea and uses evidence to demonstrate an overall understanding of the text (EALR 2, Component 1)
- Summarizes by providing an overarching statement about the text that connects to at least three events from the beginning, middle and end of text (EALR 2, Component 1)
- Infers and/or predicts about key elements of the text making connections with evidence (EALR 2, Component 1)
- Explains key vocabulary with both denotative and connotative definitions by linking them to the text (EALR 1, Component 2)

Analysis, interpretation, & synthesis of literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text

- Applies knowledge of key literary/informational elements to enhance and expand understanding of text (EALR 2, Component 2)
- Compares and contrasts ideas to explain concepts within or between text (EALR 2, Component 3)
- Analyzes text to explain the relationship between cause(s) and effect(s) and links it back to the theme or main idea (EALR 2, Component 2)

Thinks critically about literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text

- Evaluate author's/ text's purpose and/or in order to judge effectiveness on intended audience
- Evaluates reasoning of ideas / themes within the text and makes connections with evidence
- Synthesizes information beyond the text by making generalizations, drawing conclusions, or applying information to evaluate a new text or context

Table 2d (Continued)

Mathematics Criteria

Uses high school content knowledge and procedures (EALR 1) with supporting work in:

- Number Sense (EALR 1, Component 1)
- Measurement (EALR 1, Component 2)
- Geometric Sense (EALR 1, Component 3)
- Probability & Statistics (EALR 1, Component 4)
- Algebraic Sense (EALR 1, Component 5)

Solves Problems (EALR 2)

- Applies one or more strategies that lead to the answer (EALR 2, Component 2)
- Determines the answer to the problem (EALR 2, Component 3)

Reasons Logically (EALR 3)

- Justifies conclusions, results, and/or answers by addressing the conditions and/or constraints in the problem

Communicates Understanding (EALR 4)

- Gathers, represents, and/or shares mathematical information using clear mathematical language and organization

Makes Connections (EALR 5)

- Uses and relates different mathematical models and representations of the same situation using clear mathematical language and organization (EALR 5, Components 1 and 2)

Table 2e

Range-Finding – The process of selecting exemplary collections to represent different performance levels

All Content Areas

Steps in the range-finding process

- Select a range of collections to serve as potential anchors for the rubrics during scoring training, practice collections to be used for practice during scoring training, and validity collections to be randomly inserted into scoring process to ensure adherence to scoring rubrics over time
- Ensure that all selected collections have met sufficiency criteria
- Discuss scoring rubrics
- Apply scoring rubrics to selected collections
- Discuss applied scores
- Adjust scoring rubrics and/or scores, if needed, based on collections
- Assign final scores to anchor collections
- Assign final scores to practice collections
- Assign final scores to validity collections

Table 2f

Scoring Training – The process of training scorers to apply scoring rubrics consistently using anchor collections to anchor rubrics

All Content Areas

Steps in the training process

- Review and discuss rubrics
- Review and discuss anchor collections
- Score practice collections
- Discuss assigned scores; work toward consensus with pre-assigned scores
- Score second practice collections
- Discuss assigned scores; work toward consensus with pre-assigned scores

Table 2g **Table Scoring Process – The process of assigning scores to collections**
All Content Areas
Steps in the scoring process ➢ Scorers assign scores ➢ Randomly selected collections are rescored by a second scorer (inter-rater agreement) ➢ Randomly selected collections are rescored by a table leader (supervisor) ➢ Validity collections are given to scorers randomly ➢ Scorers who drift from scoring rubrics are retrained as necessary

The next two tables, Tables 3a and 3b, link each of the validity and reliability standards shown in Table 1 to the COE design features shown in Table 2.

Not all standards for the validity and reliability of classroom-based assessments can be equally addressed by design features of the COE. The most relevant issues are:

Validity Standard 6 (Consequences) requires ongoing research related to validity standards 1-5. Positive or negative consequences attributable to inferences and actions based on scores on the collections are relevant to validity ONLY if these consequences are due to weaknesses related to validity standards 1-5;

Validity Standard 4 (Curricular Validity) cannot be completely evaluated without specific information about what was taught in the courses relevant to the student work samples presented in the collections; although it is possible to minimize bias (Validity Standard) through careful selection of collections to use for scorer training, it is difficult to thoroughly assess Validity Standard 5 without more information about the students; and

Reliability Standard 2 (Clarity of Expectations) can only be evaluated if directions for assignments are provided along with students' work samples.

Table 3a: Design Features of COE that Address Validity Standards

Validity Standard	Feature of COE Addressing Std
Validity Std 1: **Authenticity**	• Protocols for R, M, W • Sufficiency Review • Scoring Rules • Range-finding • Scoring Training • Scoring Process
Validity Std 2: **Cognitive Process Validity**	• Directions for Student Assignments
Validity Std 3: **Consistency with other Measures**	• Range-finding
Validity Std 4: **Curricular Validity**	• Range-finding
Validity Std 5: **Minimizing Bias**	• Scoring Training • Scoring Process
Validity Std 6: **Consequences**	

Table 3b: Reliability Standards for Classroom-based Assessments

Reliability Standard	Feature of COE Addressing Std
Reliability Std 1: **Sufficiency**	• Protocols for R, M, W • Sufficiency Review
Reliability Std 2: **Clarity of Expectations**	• Directions for Student Assignments
Reliability Std 3: **Scoring Quality**	• Scoring Rules • Range-finding • Scoring Training • Scoring Process

Appendix A

Professional Standards for Reliability and Validity of Classroom-Based Assessments

The validity standards referenced here are derived from the work of Taylor & Nolen (1996). The original standards were adapted from the 'half dozen or so' ways of gathering evidence for the validity of the interpretation of test scores presented by Messick (1989). The reliability standards presented here are also derived from Taylor and Nolen (1996). These standards are an elaboration on three topics found in the literature on reliability of scores: a) standardization of directions (Cronbach, 1971), b) interjudge agreement, and c) internal consistency.

Throughout the literature, the term *assessment* is used to describe assessment tools (e.g., individual test questions, entire tests or quizzes, essay directions, and scoring rubrics), assessment processes (e.g., using a scoring rubric to assign points to students' essays or selecting the information to be used when assigning grades), assessment decisions (e.g., giving course grades, placing students in special programs) and assessment events (e.g., completing a test, writing a research paper, doing a course project). In the following discussion of professional standards for reliability and validity for classroom-based assessments, an attempt is made to identify these different aspects of assessment.

Validity Standards for Classroom-Based Assessments

Validity refers to the degree to which we can make accurate inferences about what examinees know and are able to do from their performances on tests and other assessments. In the classroom, validity encompasses (a) whether the assessment tools actually require students to demonstrate the knowledge and/or skills[2] described in the learning targets, (b) whether

[2] For this document ***knowledge*** includes ideas, principles, and facts as well as *understanding* of concepts, interrelationships among ideas, principles, and facts, and knowing how and when to use ideas, concepts, principles in relevant situations; ***skills*** include thinking and reasoning skills (e.g., making

instruction has prepared students for the assessed knowledge and skills *and* for the way the knowledge, understanding, and skills are assessed, (c) whether the assessment tools or processes are biased in favor of or against individuals or groups, and (d) what occurs as a result of assessment processes, events and decisions, including the full range of outcomes from feedback, grading, and placement to students' self-concepts and behaviors to students' constructions about the subject disciplines. Teachers must look at assessment tools, processes, events, and decisions for evidence of their validity. Teachers must consider alternate explanations of student performances (such as invalidity in assessments). Finally, teachers should consider the potential consequences of their assessment choices.

Validity Standard 1: Authenticity - Do the knowledge and skills required by the assessment tools reflect the way people in the subject area or discipline actually use their knowledge and skills? Before one can evaluate authenticity, one must think clearly about the domains and/or disciplines that are the focus of education and define clear learning targets related to those domains and/or disciplines. Learning targets may include knowledge and skills; learning targets may also include valued performances that require application of knowledge and skills. Sometimes teachers define their own learning targets; sometimes learning targets are provided by schools, districts, or states. With clear learning targets, the first aspect of the validity of assessments can be evaluated: Whether the assessment tool is asking students to demonstrate valued knowledge and skills in a manner that is authentic to the domain and/or discipline.

Since assessment tools also include rules for assigning points or grades, validity standard 1

---

inferences, comparing and contrasting information, drawing conclusions), research skills (e.g., skill in using the library and Internet to gather information), process skills (e.g., skill in conducting a scientific investigation, skill in using a process to go from initial ideas to a polished piece of writing), problem-solving skills, social skills, and communication skills.

also has to do with the degree to which the scoring rules and processes used to assign points or grades are tied to the learning targets *and* whether these scoring rules and processes adequately represent the domain and/or discipline. For example, effective writing involves appropriate content, relevant ideas, logical organization, word choices, language usage, appropriate voice, and writing conventions (grammar, punctuation, spelling, and capitalization); therefore, a focus *only* on writing conventions would make the assessment of writing less authentic.

Validity Standard 2: Cognitive Process Validity - Do assessments actually draw out the targeted knowledge and skills? An important aspect of validity for classroom-based assessments is whether the assessments actually require students to use the targeted knowledge and/or skills to complete a test or other performance. For example, a student might get the right answer to a multiple-choice math question because she did the same problem for homework and she remembered the answer. Another student might get the right answer because three of the four answer choices were obviously wrong. A third student might get the right answer because he copied another student's answer. A fourth student might get the right answer because he worked out the answer during the test.

Standardized test makers use "tryouts" to find out how the test questions function *before* they use the questions on tests. Most teachers do not have the luxury to do this with their own assessments. Textbook assessments are rarely tried out with students before they are published. Therefore, teachers need to develop ways to find out whether test questions and performance directions actually tap into the concepts and skills they are intended to assess.

One way to do this is to ask students to explain their work or show their steps as they complete various assessments. When students explain their reasoning, their choices, and their

solutions, a teacher may discover that an assessment isn't really tapping into the targeted knowledge and skills. If this is the case, test questions and performance directions can be adjusted to ensure that students must use / demonstrate the targeted knowledge and skills in completing the assessment.

Validity Standard 3: Consistency with other Measures - Do students perform similarly on different measures of the same knowledge and skills? Another aspect of validity is whether students perform similarly on assessment tools that are intended to measure the same learning targets. One strategy for determining whether assessment tools and processes can be used to make valid inferences about examinees is to have students do more than one version of the same type of work. For example, a teacher might have 3-4 questions on a test to assess a particular science concept. She might have students do 2-3 science investigations to assess students understanding of investigative procedures. Multiple pieces of evidence provide information about whether students are performing similarly on assessments that are intended to measure the same thing. In short, for Validity Standard 3 teachers can review several sources of evidence to see whether examinees perform consistently across different assessments of the same knowledge and/or skills. If student performances on different tasks measuring the same knowledge or skill are very similar, the teacher can have more confidence that the test questions or performance tasks are measures of the same learning targets. The grade record in Figure 1 shows student performance on six essays, all of which were evaluated with the same two scoring rubrics – a five point rubric for content and a five point rubric for writing conventions. As can be seen, despite the fact that several students in the class have earned scores of 5, the highest score for Essay 4 was 3. This suggests that there may be a problem with the validity of scores for Essay 4.

Figure 1

An Example of Inconsistency of Assessment Scores as a Potential Threat to Validity

STUDENT SCORES ON 6 ESSAYS

	Essay 1		Essay 2		Essay 3		Essay 4		Essay 5		Essay 6	
Student	Cont.	Conv.	Cont.	Conv.	Cont.	Conv.	Cont.	Conv.	Cont.	Conv.	Cont.	Conv.
Tanya	5	5	5	5	5	5	3	5	5	5	5	5
Mario	5	5	5	5	5	5	3	5	5	5	5	5
Emma	2	3	2	2	4	3	3	4	4	4	5	5
Juan	3	3	4	3	4	4	2	4	4	4	5	4
Geoff	2	3	3	3	3	3	3	3	3	4	4	4
Robin	4	5	4	5	4	5	3	5	5	5	5	5
Caitlyn	5	5	5	5	5	5	3	5	5	5	5	5
Points Possible	5	5	5	5	5	5	5	5	5	5	5	5

Cont. = Content
Conv. = Writing Conventions

Validity Standard 4: Curricular Validity - Does the assessment fit the content taught and the instructional methods used? One of the most fundamental validity questions a teacher should ask is whether the learning targets were actually taught, whether the method of assessment fits the *way* knowledge and skills were taught, and whether students had sufficient exposure to and practice with knowledge and skills to be successful on the assessments. For example, if students are asked to practice routine mathematical algorithms in class and for homework but are then asked to apply the algorithms in novel situations on a test, the assessment tool is not valid for the instructional context. A mismatch between what is taught and what is assessed can lead to frustration for teachers and students and can also result in invalid grades for students.

Validity Standard 5: Minimizing Bias - Does the assessment

tool, process, and/or event work equally well across individuals, groups, settings? In addition to fit with instruction, validity has to do with how well various assessment tools, processes, and/or events allow all students to demonstrate their knowledge and skill. When an assessment favors some students over others, this is called bias. Bias occurs whenever students who have achieved the valued knowledge and skills do not or cannot demonstrate their achievements because of some aspect of the assessment tool. Teachers need to know whether differences in performance across students are because of true differences in students' knowledge and skills or whether differences are due to invalidity in the assessment tools, processes, or events.

Assessment decisions may be invalid when factors *within* the assessment tool prevent students from showing what they know and are able to do. One factor that might affect students performance is when the *context* or *content* is unfamiliar to students *and* unrelated to the learning targets. For example, an assessment might require students to write on a topic about which they have little or no experience (e.g., "Write a story describing something that happened at Thanksgiving dinner."). Although students might be able to write effectively, the writing topic may prevent some students from demonstrating their writing skills. An important distinction here is that the context of the writing prompt is unrelated to what the teacher wants to know – whether students can write using important knowledge and skills related to the characteristics of effective writing (e.g., organization, word choices), the writing purpose (narrative), and writing conventions (e.g., grammar and spelling). When the context set for an assessment favors some students over others, the assessment tool is biased. Teachers are responsible for creating assessment contexts that allow *all* students to demonstrate their knowledge and skills. This may mean that the contexts are different for different students. For example, the writing teacher could

provide different writing prompts and allow students to select the one that works best for their backgrounds.

A related source of potential bias is whether the *format* of the assessment tool prevents some students from demonstrating their knowledge and skills. Suppose a teacher wants to assess students' understanding of character development, plot development, theme, and setting. He assigns the same novel to all of the students in his class, and asks for a written essay. To demonstrate their literary analysis skills, students must read and write. Suppose, also, that some students are English language learners (ELL) who have good literary analysis skills but cannot read the novel because it is too difficult and cannot write the essay because they are not yet skilled writers. A different assessment format may be required for these students (e.g., hearing a book on tape and giving an oral report) in order to make valid inferences about their literary analysis skills.

Similar questions can be asked about any assessment. In creating and selecting assessments, teachers must determine whether student work is influenced by factors irrelevant to the targeted learning objectives such as assessment context, format, response mode, cultural experiences, or other factors. If the learned knowledge and skills *can* be demonstrated in a way other than through a specific assessment tool (without changing the target for what is assessed) and if some students can show their knowledge, conceptual understanding and skills through the alternate format, then a single format for the assessment tool is *biased* in favor of those who can perform in the chosen way and against those who cannot.

A third potential source of bias comes from the rules used to assign points to students work. To be valid scoring rules, the rules must focus *only* on the targeted knowledge and skills. For example, suppose a teacher evaluates literary analysis essays are using scoring rules that award

points based on elegance of the writing or the creativity of presentation rather than the adequacy of literary interpretations. The assessment process will be biased in favor of students who are skilled or creative writers.

Finally, another source of potential invalidity related to standard 5 comes from the teacher. If the teacher 'colors' the process used to assign points to students' work with prior knowledge of students or attitudes toward students – rather than consistently applying scoring rules across all students' work – the resulting scores may not be valid reflections of students' knowledge and skills.

Validity standard 5 becomes increasingly critical as classrooms become more diverse and whole-group teaching becomes more difficult. Teachers must provide appropriate adaptations of assessment tools and processes while still obtaining valid evidence about student achievement related to the learning targets.

Validity Standard 6: Consequences - Are there negative consequences that could be prevented if assessment tools, processes, events, or decisions had been more valid? Assessments tools, processes, events, and decisions have effects on students. Tests, projects, teacher feedback, and grades can all influence student learning, self-concepts, motivation (Butler & Nisan, 1986; Covington & Omelich, 1984), and perceptions of the subject areas and disciplines being taught. Therefore, the final standard of validity for classroom-based assessments is related to how classroom assessments affect the students themselves. If students develop a notion of the discipline of history as a collection of facts that are to be memorized, this consequence is *mis-*educative. If some students get poor grades whereas others get good grades *because of invalidity in standards 1 through 5,* then the consequences that arise from those grades (promotion to the

next grade level, placement in special programs, access to honors classes, etc.) are invalid consequences. Educators have an ethical responsibility to create and select valid assessment tools and to use valid processes so that consequences are fair, are based on appropriate information, and do not create misconceptions for students.

Reliability Standards for Classroom-Based Assessments

Reliability in classroom-based assessment refers to the degree to which one can rely on the results of assessment processes and events. Three standards of reliability are relevant to the classroom: (1) whether one can make dependable statements about what students have learned in relation to the learning targets, (2) whether students know exactly what is expected on tests, performances, and other assessment events so that their performance is consistent with their level of knowledge and skill, (3) whether the scoring rules and assessment processes are systematic enough to ensure that evaluators are consistent across students and over time.

Reliability Standard 1: Sufficiency - Is there sufficient evidence of student learning so that one can make a dependable statements about what each student has learned in relation to the learning targets? Assessment experts often talk about reliability as consistency in performance. Is a single throw of the toy basketball sufficient to make a dependable (reliable) statement about whether the student could throw a ball into a basket? Would the student perform in the same way a second time? To make a reliable statement about what students know and are able to do, we can ask them to do similar tasks several times (e.g., reading tests, mathematics problems) and look to see whether their performance is consistent over time. Summative

decisions made at the end of a grading period[3] can be much more reliable than the results of individual assessments. For summative decisions to be reliable, one must ensure that there is *sufficient, high-quality* assessment information from which to make reliable decisions about students. The reliability of summative decisions depends on the validity of the assessment tools and processes. If attention is given to validity standards one through five, then one can begin to ask whether there is sufficient information from which to make reliable decisions. Multiple, valid assessments are very likely to give reliable information about students. The more sources of valid assessment information teachers have at the end of a grading period, the more likely that their decisions will be ones that they and others can trust. Therefore, to address Reliability Standard 1, one must obtain as much valid information about students' achievement of the learning targets as possible. Classroom teachers can and should bring a wide range of information – observations, test scores, homework, class work, written papers, etc. – to bear on summative decisions such as course grades.

Reliability Standard 2: Clarity of Expectations - Do the test items or performance directions provide clear, unambiguous expectations for students? When students are not clear about what they are being asked to do, they are less likely to produce the expected response; they are more likely to respond in a way that is *inconsistent* with their own knowledge and skills or in ways that are inconsistent across tasks for which you have the same expectations. In contrast, when test items and performance directions are clear and explicit, students are more able to show what they know and are able to do. When directions and items are clear and focused, the quality of student work is likely to be much better and easier to evaluate. When students are clear about what they

[3] A grading period is the time between report cards such as a quarter, trimester, or semester.

are to do, their performances and responses are more likely to reflect their true knowledge and skills.

Reliability Standard 3: Scoring Quality - Are the scoring rules and evaluation processes systematic enough to ensure consistent evaluation across students and over time? There are generally three types of assessment tools that could be affected by the consistency of judgments about students' learning: short-answer and performance questions for tests; projects and performances; and multiple assignments for which a teacher has the same expectations. In these three situations, the consistency of judgments depends on (a) whether the rules for scoring short-answer items and performance items are specific and clear enough that they can be applied consistently across students, (b) whether the rules for scoring extended performances are specific and clear enough that they can be applied consistently across students, and (c) whether rules for scoring frequently occurring assessments are applied consistently across similar tasks and over time.